DOCUMENT RESUME

ED 152 332 IR 005 914

AUTHOR Tessier, Judith A.; And Others
TITLE Feasibility of Maintaining and Providing Access to Data Archives through ERIC.
SPONS AGENCY Syracuse Univ., N.Y. ERIC Clearinghouse on Information Resources.
PUB DATE Dec 77
GRANT NIE-R-76-0027
NOTE 79p.; For related document, see IR 005 915

EDRS PRICE MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS Archives; Data; *Feasibility Studies; *Information Networks; Information Science; *Information Services; Instructional Technology; Library Science; Shared Services; *Statistical Analysis; *Use Studies
IDENTIFIERS *ERIC; Information Analysis Products

ABSTRACT

This study explored the feasibility of establishing archive services in ERIC/IR (ERIC Clearinghouse on Information Resources) to make available survey data for new studies or secondary analysis by researchers in the areas of library and information science and instructional technology. A community of potential users and producers was identified and surveyed to determine both availability of data and demand for sharing data sets, as well as types of data, attitudes of researchers toward sharing their data, and formats in which data are available. Eight alternative levels of data services were identified, ranging from continuation of the current system of informal exchange to an archive with such added services as statistical summaries and data manipulation. These alternatives were considered in terms of adaptability to the ERIC system and cost effectiveness. Findings indicate that a data sharing service is needed, and that ERIC/IR is an appropriate service center. The level of service recommended, in view of costs and amount of data available in other than machine-readable formats, is the listing of data sets in Resources in Education (RIE) with information on availability, and publication by ERIC/IR of an annual directory of these data sets. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152332

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

FEASIBILITY OF MAINTAINING AND PROVIDING ACCESS TO DATA ARCHIVES THROUGH ERIC

Judith A. Tessier

Jennifer Kuehn

Terry Noreault

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Judith A. Tessier

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Report of a Special Project cponsored by the ERIC Clearinghouse on Information Resources and the School of Information Studies, Syracuse University, Syracuse, New York.

IR005914

ERIC Clearinghouse on Information Resources
Syracuse University
December 1977

Judith A. Tessier is a lecturer in the School of Information Studies, Syracuse University.

Jennifer Kuehn and Terry Noreault are doctoral students in the School of Information Studies, Syracuse University.

The material in this publication is based upon research supported in part by the National Institute of Education, United States Department of Health, Education, and Welfare under Grant No. NIE-R-76-0027 and by the School of Information Studies, Syracuse University, Syracuse, New York. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Institute of Education or the School of Information Studies.

ABSTRACT

This study explored the feasibility of establishing archive services in ERIC/IR (ERIC Clearinghouse on Information Resources) to make available survey data for new studies or secondary analysis by researchers in the areas of library and information science and instructional technology.

A community of potential users and producers was identified and surveyed to determine both availability of data and demand for sharing data sets, as well as types of data, attitudes of researchers toward sharing their data, and formats in which data are available.

Eight alternative levels of data services were identified, ranging from continuation of the current system of informal exchange to an archive with such added services as statistical summaries and data manipulation. These alternatives were considered in terms of adaptability to the ERIC system and cost effectiveness.

Findings indicate that a data sharing service is needed, and that ERIC/IR is an appropriate service center. The level of service recommended, in view of costs and amount of data available in other than machine-readable formats, is the listing of data sets in Resources in Education (RIE) with information on availability, and publication by ERIC/IR of an annual directory of these data sets.

ACKNOWLEDGMENTS

We would like to express our appreciation for the support we received from Professors Pauline A. Atherton of the School of Information Studies and Donald P. Ely and Richard E. Clark of the Area of Instructional Technology, Syracuse University. Matthew Koll, a doctoral student in the School of Information Studies, made major contributions to the cost section and provided day-to-day support on the survey. We would also like to thank the many people who corresponded with us and those that took the time to provide information or answer our survey.

TABLE OF CONTENTS

	Page
ABSTRACT	1
ACKNOWLEDGMENTS	ii
INTRODUCTION	1
Definitions	2
Background	3
Statement of research problem	6
Description of alternative levels of data services	8
METHODOLOGY AND FINDINGS	13
Availability of data: Findings	13
Methodology	13
Estimate of data availability	14
Characteristics of available data	16
Diversity of data sets	16
Researchers' willingness to share data sets	17
Popularity of data sets	19
Machine-readability of data sets	20
Estimate of data sets available to ERIC	21
Demand for data sets: Findings	23
Inadequacies of current data sharing system	23
Characteristics of data sets shared	25
Uses and benefits of data sharing	27
Preferences and implications for ERIC/IR	28
Organization and description of data sets in ERIC/IR	30
The ERIC format	33
Alternative 1: Treatment as a supplement to the resumes in ERIC indexes (RIE and CIJE)	35
Availability	35
Abstract	36
Indexing	36
Alternative 2: Treatment as a separate item in ERIC	38
Copyright, reproduction release, level of availability	38
Publication type	38
Abstract and indexing terms	39
Summary	40
Cost evaluation of alternative services	42
1. No change	43
2. Listing in RIE	43
3. Data set directory	44
4. Archive	45
5. Archive with added services	47
User costs	47

Page

CONCLUSIONS AND RECOMMENDATIONS 49

APPENDIX I. Description of sample, survey, and instruments 55

APPENDIX II. Data archives from which information was received 65

APPENDIX III. ERIC document resume form 66

APPENDIX IV. Sample data set descriptions 67

REFERENCES 70

List of Tables

Table 1. Alternative data service levels and functions 11

Table 2. Samples of data sets generated 15

Table 3. Generators' willingness to share data sets 18

Table 4. Format for sharing data 20

Table 5. Listing of data sets shared 26

Table 6. Purposes for secondary use of shared data sets 27

Table 7. Results of use of shared data sets 27

Table 8. Data elements required for inventory purposes 30

Table 9. Applicability of ERIC description to data sets 34

Table A-1. Response rate 62

Table A-2. Postcard responses. 63

Table A-3. Access to computer and facilities 64

Table A-4. Experience with computers 64

INTRODUCTION

Research is performed but rarely replicated. Research data are collected but rarely re-analyzed. Research problems are studied year after year without ever using data from similar previous studies. Realizing that many of these data sets are in machine-readable form, and possibly useful and available to others, this project explored the feasibility of establishing an archive for machine-readable data within the scope of the ERIC Clearinghouse on Information Resources (see page 3). The community of potential users and producers was surveyed to estimate demand for archive services and availability of data. Factors of archive management within the ERIC system were also studied. Conclusions and recommendations are based on cost estimates and potential benefits.

In an effort to consider the issues concisely, the study method and results have been summarized within the body of the report. Full descriptions of the study sample and the response rate are appended to this volume: a display of survey instruments and data is available as a separate ERIC document.

Major recommendations are for the inclusion of data availability information within the ERIC system, and publication of a directory on a continuing basis. An archive is not recommended, as this is a costly undertaking, given the data potentially available and the level of expressed demand.

Definitions

Because survey data archives originated in the social sciences, the terminology used in describing functions, problems, and techniques reflects a social science perspective. In considering an archive service for information resources, it became apparent that definitions pertinent to the potential service audience were needed.

The definitions noted here were developed during the course of the project. They were used to guide enquiries, letters to respondents, questionnaire construction, and analysis.

DATA: numerical description of survey or experimental cases generated in the process of doing research. Cases might be people, institutions, interactions, classes, document queries, or other units of analysis. This definition differs somewhat from social science archives where survey data alone is typically controlled (assumed to be numerical and excluding experimental studies). In the ERIC Clearinghouse on Information Resources (ERIC/IR) scope, experimental studies are included. While this study places primary emphasis on numerical data, some non-numerical data have been identified, i.e., videotapes. Bibliographic data are excluded.
Data set is used here to designate the logical file of data as seen by the researcher. Typically a data set is the entire body of raw data organized so that cases and variables are accessible. A data set may contain the results of some data transformations, but does not include tables displaying only summary statistics.

MACHINE-READABLE DATA: data readable by computers or electronic data processing equipment so that secondary use does not require manual input. Typical storage media are punched cards and magnetic tape. Videotaped case studies (not numerical descriptions and requiring manual input) are not included in the definition. Photocopies of data collection forms or code sheets are also excluded.

DATA ARCHIVE: a service center which gathers, preserves, and disseminates data for secondary analysis. Typically archives include the functions of acquisition, description, indexing, inventory control, cleaning and duplication, and may also provide codebooks and other documentation and allied services such as statistical consultation, instruction, or analysis.

DATA DIRECTORY: a bibliography or listing of data sets that researchers have agreed to duplicate for secondary analysis; the listing includes source information, data set description, and constraints on use. The term is used here to distinguish between a listing of available data sets and a centralized data archive.

ERIC: the acronym for the Educational Resources Information Center, a service system designed to provide bibliographic control and access to the literature in education for the education community. The system is made up of a national center and facility for administration and coordination, a set of sixteen Clearinghouses for special topics in education, and agreements with private companies for the publication of its products. Two major publications providing access to the literature are: Resources in Education (RIE) for report literature and Current Information in Journals of Education (CIJE).

INFORMATION RESOURCES (IR) CLEARINGHOUSE: The ERIC Clearinghouse for the special topic "Information Resources," encompassing instructional technology, library science, and information science, located at Syracuse University since January 1, 1977. The scope of this Clearinghouse includes the acquisition and processing of documents on research, development, delivery, and evaluation of information and instructional technology. It focuses on personnel, personnel development, strategies, systems, procedures, materials, and equipment used in the fields of information and education. Current areas include libraries; learning resource centers; information science; instructional design, development, and evaluation; systems analysis; community information systems; and instructional media, as well as strategies which flow from these topics: e.g., information transfer, mastery learning, simulation and gaming, programmed instruction, and information brokering. The Clearinghouse is also concerned with the delivery of information and instruction through media such as television, computers, radio, films, microforms, holography, and other audiovisual devices.

Background

The fields within the scope of the ERIC/IR Clearinghouse--instructional technology, library science, and information science--have experienced that same data explosion so often cited in the hard sciences and the social sciences. Quantitative research--generating data--has increased, primed by federal funding and by the application of social

science methodologies. Moreover, the recognition of the usefulness of continuous reporting statistics for federal program planning in education and libraries has resulted in another kind of data increase.

Information resources research today is analogous to the state in social science research ten years ago. Nasatir (1973) cited the conditions: increased governmental statistics gathering, increased polling and market research, more data gathered by the individual researcher than he will use. He notes:

> The result has been a torrent of data from these sources which in many cases has paralysed and immobilized researchers and decision makers rather than assisting them in their quest for knowledge. This paradox arises from the recognition that the desired data may already exist--somewhere, but is not readily available. (p. 8)

Ratnofsky and Corrallo of the Office of Planning, Budgeting and Evaluation at USOE (1977) expressed a similar concern for economy and the desire to avoid duplication of data collection. A major point of their report was a preference for already collected data if these could be found that met the requirement of the research problem.

Another historical trend should be noted. The work by Cronbach and Snow (1977), reinterpreting research on classroom treatments, and the work by Zweizig and Dervin (1977), which was an inventory and critique of information needs studies, are but two examples. In both cases, the researchers were able to look at the cumulated set of studies and draw inferences that led to a new set of research questions. The body of research is growing. But that style of research--secondary analysis and middle range theory building--rests on the ability to access and cumulate previous work.

Use of data for new studies and use for secondary analysis have proven the worth of the social science survey data archive. First the data archive provides access to individual data sets: for verification, reinterpretation, or for economy when the data can be applied to another research question. Second, a data archive cumulates data, providing further benefits through access to multiple data sets:

- estimates of bias, when two independent studies contain the same population.
- tests of generalization and robustness of theory across specialized populations.
- "testing of many complex hypotheses not one of which was ever anticipated the initial researchers and gatherers of the data."
- systematic review over historical periods.
- instruction in intricacies of data analysis. (Nasatir, p. 8)

Data resources in the field of Information Resources are currently uncontrolled. This anecdotal excerpt from Report on Education Research (June 16, 1976) states the case in education generally:

> WHATEVER HAPPENED TO THE COLEMAN DATA TAPES? Data from original research studies ought to be preserved, education evaluation expert Gene Glass told the recent annual convention of the American Educational Research Association in San Francisco. Glass, AERA's outgoing president from the University of Colorado, provided the following example:
>
> "One of our graduate students, Karl White, recently spent 15 hours and made 30 phone calls in an attempt to obtain from the government a copy of the data tapes for the Coleman study only to learn in the end that they had been irretrievably filed in unmarked tape cannisters with some 2,000 other unmarked data tapes. The tapes are available from the research underground, but the whole business is too important to be so treated." (p. 8)

Statement of the research problem

Is it feasible to establish a data archive service covering the scope of ERIC/IR?

This feasibility study arose out of a concern that data should be accessible in the fields of Information Resources. The benefits of access were cited above:

- economy through full exploitation of already collected data.
- consolidation of knowledge through critical reanalysis and cross study comparison.

The concept of an archive was proposed because it has seen success in the social sciences and because it provides efficiencies in the data sharing process. One deposit copy is required from the source, while duplication for multiple requests becomes an operation of a specialized staff.

In proposing a data archive service for Information Resources, it became clear that the availability of data that might be archived was an unknown. Similarly, the demand or potential use of data for secondary analysis was unknown. Design of a system without basic information on available data and potential use was premature. The feasibility study then required three parts:

- estimate of data availability;
- estimate of demand for data; and
- system design considerations.

Also, the literature in social science data sharing illustrated that a continuum of service levels is possible. Archiving is one alternative; however, the possibility of providing more limited access through a

directory is a less expensive type of service to be considered. A directory places more responsibility on the source of the data: he is responsible for the dissemination of data sets and no archive distribution center is necessary. The directory, essentially an alerting and source information service, was an alternative that needed to be incorporated in the feasibility study.

Availability of data, demand for data, and level of service interact. The willingness of the researcher to deposit data depends in part on the credibility, control, and rewards of the system of exchange (Miller, 1976). In similar ways, the requests for data depend on coverage, access, levels of service, and user characteristics.

The methodology used in the feasibility study reflects these considerations. First, possible levels of service were blocked out. Second, data availability was assessed by querying researchers in the Information Resources field. At the same time potential demand, evidenced through past sharing and desire to share, was explored. Potential users were also queried on preferred levels of service. These steps provided some information on availability and demand, although projections for a hypothesized information system are subject to variables only testable in operation.

The ability of the ERIC system to provide services required of an archive was considered. Particular consideration was given to acquisition and control functions. ERIC has a set of operations through which some levels of service in an archive might work. Attention was also given to the costs of data archives in the social sciences.

System characteristics and costs were considered in contrast to availability and demand. The final question of feasibility has been couched in these terms: what are the feasible levels of service, given

the characteristics of the data, the uses, the costs, and the identifiable benefits?

Description of alternative levels of data services

This section describes eight alternative levels of service and the functions of the service center at each level.

These alternatives have been identified as:

1. No new service instituted by ERIC: Continue the current system of informal exchange.

2. Listing of data sets in ERIC Resources in Education: This implies an on-going listing, and indexing through RIE but with responsibility for duplication with the data originator.

3. Data directory: An ERIC/IR bibliography or directory of data in the field that researchers offer for duplication, listing any constraints on use, etc.; again with duplication responsibilities for the originator.

4. Archive: An ERIC service center that gathers, preserves and disseminates data for secondary analysis.

5. Archive with added service of data manipulation in preparing subsets of data on demand.

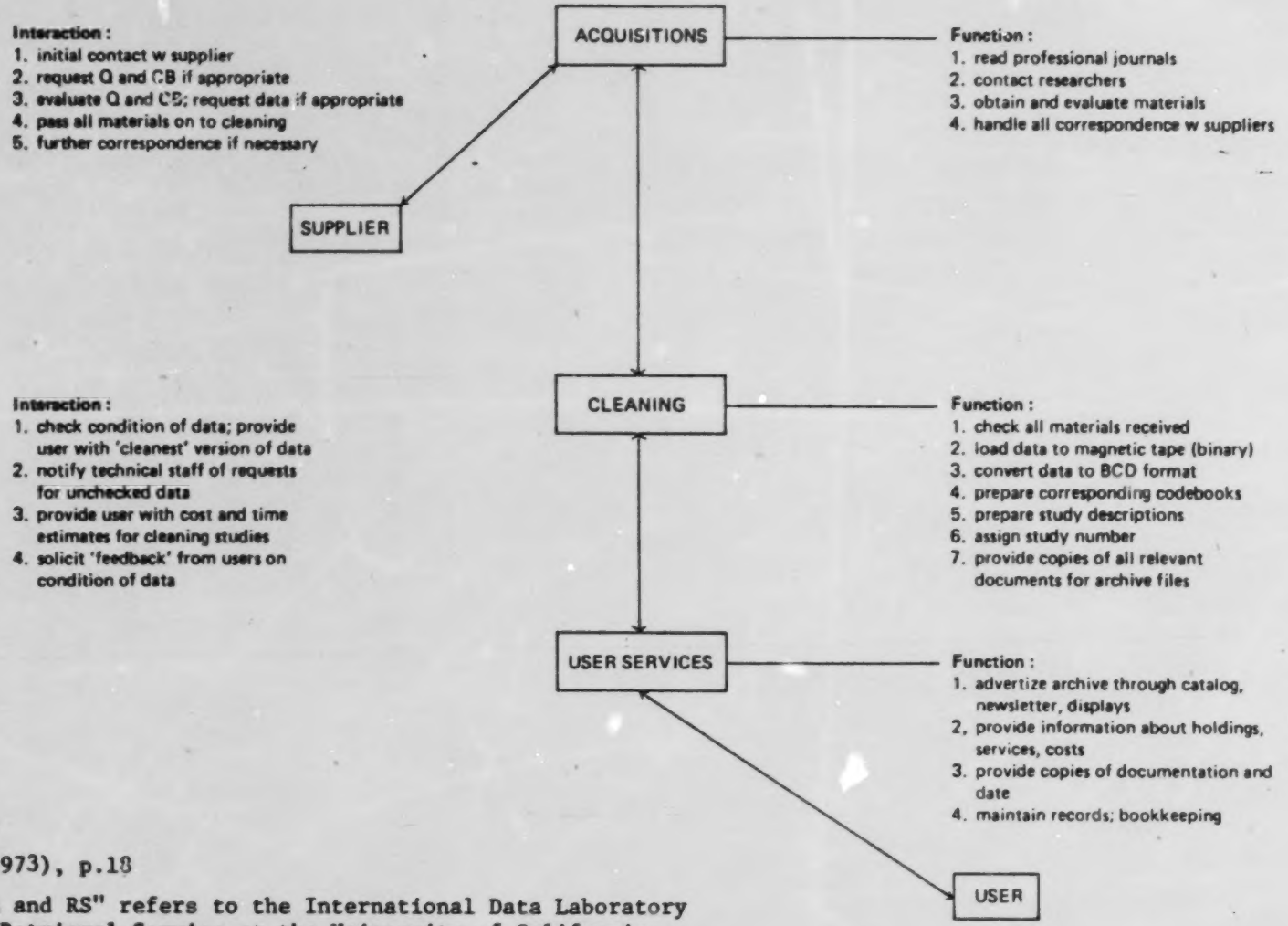
6. Archive with added service of statistical summaries or analysis and possibly cross-study comparisons.

7. Archive with statistical consulting on the use of data sets.

8. Archive with the development of instructional data sets and instructional modules for data analysis.

Publication of a data directory and listings in RIE are not archival functions, in that these do not require the holding and duplication of the data set by one centralized entity. The reasoning for inclusion was that system design is secondary to supporting the benefits of access to data. Since directory listing or RIE listing provides access (at least to finding information), these alternative services were included. Paul Peters (1974) argues:

FLOW CHART OF IDL AND RS FUNCTIONS


Interaction:
1. initial contact w supplier
2. request Q and CB if appropriate
3. evaluate Q and CB; request data if appropriate
4. pass all materials on to cleaning
5. further correspondence if necessary
ACQUISITIONS
Function:
1. read professional journals
2. contact researchers
3. obtain and evaluate materials
4. handle all correspondence w suppliers
SUPPLIER
Interaction:
1. check condition of data; provide user with 'cleanest' version of data
2. notify technical staff of requests for unchecked data
3. provide user with cost and time estimates for cleaning studies
4. solicit 'feedback' from users on condition of data
CLEANING
Function:
1. check all materials received
2. load data to magnetic tape (binary)
3. convert data to BCD format
4. prepare corresponding codebooks
5. prepare study descriptions
6. assign study number
7. provide copies of all relevant documents for archive files
USER SERVICES
Function:
1. advertize archive through catalog, newsletter, displays
2. provide information about holdings, services, costs
3. provide copies of documentation and date
4. maintain records; bookkeeping
USER

Nasatir (1973), p.18

Note: "IDL and RS" refers to the International Data Laboratory and Retrieval Service at the University of California at Berkeley.
"Q" = questionnaire; "CB" = code book.

> The ostensible goal of a social science data dissemination system is to communicate raw data for use by individuals other than those responsible for collection of the data and their primary analysis. Without ignoring this goal it is important to consider the alternative goal of disseminating information about data. The two goals are intimately related. The question is one of priorities. The opinion advanced in this paper is that the goal of disseminating information about data should be assigned the higher priority. (p. 8)

Bisco (1970) lists the following functions of an archive:

1. Information acquisition: locacing relevant data and ancillary information and arranging for their acquisition, duplication and subsequent diffusion.

2. Processing: preparation for distribution and analysis.

3. Maintenance: insuring currency and accuracy and back-up copies.

4. User services: preparing duplicates, subsets, data analysis output, consultation in mathematics, statistics, methodology, programming and social research.

5. Training: programs and research seminars.(p. 8-9)

The services listed by Bisco reflect those offered by the Inter-University Consortium for Political and Social Research (Ann Arbor, Michigan), which he headed. The archive stresses maximum use of a reasonably homogenous set of data tapes.

Nasatir (1973) uses a diagram to identify functions and limits user services to those of data set delivery (see page 9).

The distinctions in Nasatir's user services and Bisco's led the feasibility group to consider alternative levels of user services within the archive model also. The necessity of these services has been a source of contention within social science archives; prescription to another field of research was inappropriate.

The literature then suggests an array of possible services, each of which provides access to data in some way. The table below is a sketch of archive functions necessary to achieve a given level of service and the consequent benefits to the users.

Table 1. Alternative data service levels and functions

Service Functions	Service levels		
	Listing or directory	Archive	Archive with options
Acquisition	Information about: data codebooks questionnaires source	Data Codebooks Questionnaires Releases or constraints	Same as archive
Processing	Study description	Study description Cleaning (level varies by archive requirements)	Same as archive and Programming for local system for each option
User services	Catalog entries	Data delivery Backup information Usage constraints	Same as archive and subsets statistics consulting instruction
Benefits	Awareness service Access to data through researcher Maintenance of informal communication	Awareness Access Delivery Efficiency for researcher	Same as archive ease of use for cross study comparison

The level of service provided will affect what data can be made available and uses of the service. Bisco argues that an archive that provides data delivery and not allied services fails to meet its user's demands. It can also be argued that demand increases as level of service is increased. This may be particularly pertinent where the service audience is not sophisticated in data handling and statistical analysis. A design question then is whether use is dependent on some minimal level of service. The next two sections identify data availability and demand characteristics required to identify the appropriate level of service in the Information Resources field.

METHODOLOGY AND FINDINGS

AVAILABILITY OF DATA: FINDINGS

This section describes our findings on the availability of data sets in the ERIC/IR Clearinghouse scope. The production of data sets, their availability for sharing, and their characteristics are considered.

Data are available for data sharing; some are available in machine-readable form. The conclusion is supported both by the holdings and data generating activities of our respondents and by the expressed willingness to share through a directory or archive.

Methodology

The availability of data and the demand for a data sharing service were explored through a survey to potential contributors and users. The sample design, survey administration, response rates, questionnaires, and results are described in the appendices. The following brief description is included here to facilitate the reading of the main report only.

A sample of contributors to CIJE and RIE (n = 114) and educators in instructional technology, library science, and information science (n = 112) were surveyed with a postal card return on general data sharing activities. The respondents to the first mailing of 226 (n = 114, or 50%) then received a second customized questionnaire (response rate: 46%) based on their first responses. Four major categories were used to differentiate respondents:

- used someone else's data.
- wanted to borrow someone else's data, but did not.
- generated data which might be of interest to others.
- shared your data through any means (informal borrowing, archives, etc.).

Special sections were also sent to doctoral advisors and instructors of research methodology.

Estimates of data availability

A study conducted for survey data in Great Britain (Phillipson, 1968) indicated some of the problems encountered in estimating availability: identifying likely sources; requiring respondents to list data sets perhaps inventoried only in the common knowledge of the organization; considerations of propriety when the system of exchange is unstated. Two techniques were used to estimate availability here: a description of the data generating activity of the respondents and a projection of possible data available for ERIC/IR reports in RIE and CIJE.

Our respondents, identified as researchers by their contribution to ERIC/IR or as teachers in research, identified themselves as researchers (71%) and as having generated research data of interest to others (76%). When asked about their data on the second wave, generators indicated a diversity of number of data sets generated in the past two years, but overall, averaged 1.3 data sets a year. For the whole population (over 1200, 75% generators, at 1.3 yearly) the projection of productivity is well over 1000 data sets generated per year. This projection seems highly inflated and may be applicable only to a smaller subset of highly active researchers. The findings indicate (1) that productivity varies from a periodicity of more than 2 years to one researcher who generated 10-20 data sets in the two year period (the average for respondents was 1.3 per year); and (2) that most recent named data sets (possibly those deemed more appropriate for sharing) varied in dates from 1971 to 1977 (table 2). It is also interesting to note that researchers listed fewer data sets than they claimed to have generated. These conditions suggest that an estimate of productivity from researchers is subject to a variety of uncontrolled variables.

Table 2. Sample of data sets generated

Data set	date	Data set	date
Acquisition of Comprehension	76	Information needs of Blue Collar Adults	75
Acquisition of Comprehension: reading material	77	Library Circulation	71
Acquisition of relative clauses	75	Linguistic analysis of Sci-Tech	75
All Library Statistics	54-76	Museum Libraries	n.d.
Apprehension among student teachers	77	Performance effect of prize probability	77
Approval plans	n.d.	Pilot Test for Library Statistics data base	77
Attitude change	n.d.	Pretest Cognitive Achievement on test	75
Ballots Users	77	Update	75
Book models & Paperbacks	75	Update	76
Update of above	77	Self-citation in articles in plant physiology and neurobiology	77
Catalog use	75	74/75 analysis of Library Statistics	76
Censoring: including periodical	75	Statistics of Public Libraries over 100,000	73
Cost data for technical process	n.d.	Update	75
Comprehension of teaching methods to nursing students	72	Update	77
Cost-Effective Book pick-up	71	Statistics US Libraries 1876	53-77
Decision making	n.d.	Survey of Costs in Technical Processing & Interlibrary	76
Demography of University population	77	Survey of Library Use by Studies in External Degtee Program	75
Drug use	77	Warranted Uncertainty	76
Effect of prize on P/A tasks	76	Women information needs and uses	74
Index of State Library Statistics	76		

Underlined data sets have been shared.

The second technique for projecting availability started with our sample selection for RIE and CIJE contributors. In the selection of authors, an estimate of IR reports for which data might exist was made (see survey sample, appendix). Thirteen percent of the CIJE items and 25% of the RIE documents appeared to be based on research data. For the two-year sample period, just over 1000 items were identified, or an annual average of 500.

The marked difference in the two estimates arises first from the ERIC/IR screening process, through which only documents meeting scope and criteria are included. The existence of a document based on a data set requires a two-step process and to some degree a reduction in data sets from those generated to those used. Finally, the contributions to ERIC and the contributors are only a subset of the research and researchers in the field for any given time period. These conditions taken together help to account for the differences in estimates.

Characteristics of available data

(1) Diversity of data sets: The diversity of the data generated is indicated by a partial listing of holdings cited by respondents (table 2).

At least two respondents excluded themselves from the second questionnaire on the grounds that they had misunderstood our use of the term data. One case concerned video tapes. The second was less clear-cut. The respondent and his work were known to the feasibility project staff and had been identified as eligible, but the researcher excluded himself on the grounds that he had "no data"--probably because much of the research is based on bibliographic data. The diversity of data in the Information Resources scope distinguishes this effort from survey archives and creates some unique problems in establishing what is eligible.

Responses to the request for a data listing for an ERIC item predefines "appropriate data" in a way the questionnaire could not. First,

if ERIC seeks data sets, the criteria for identification from the controlled reports requires (1) that the data have been used, and (2) that the data are within the ERIC/IR scope. Moreover, the researcher's response to queries for data used in publication can be expected to be more specific than the researchers' responses to the survey query: What data have you generated? For these reasons, the estimate from the CIJE/RIE sample is preferred to the activity reported by data generators when considering available data sets for a sharing service through ERIC. This more conservative estimate is used in the cost section of this report.

(2) Researchers' willingness to share data: Only a subset of those data sets cited or implied are available and appropriate for a data sharing service. Miller (1976) and Bisco (1970) cite reasons that good data sets are not controlled by data sharing services: unwillingness of the researcher to share his data and technical limitations such as machine-readability, multiple punches, absence of documentation, and the like.

The researchers in our sample, by and large, were willing to share their data. Those who have shared data indicated their willingness to continue to share. Generally data sharing was not considered on a cost-recovery or profit making basis: it was viewed as a professional courtesy. One person stated: "My remuneration was in the satisfaction of helping others."

Most generators of research indicated willingness to share data sets (table 3).

Rather than suggest barriers to data sharing, the questionnaire asked researchers to state conditions that they felt would affect their decisions to list or archive. These comments are informative and echo Miller's (1976) note that the researcher wants to reduce his own load in data sharing,

Table 3. Generator's willingness to share data sets

	To list in directory		To archive	
	N	(%)	N	(%)
Yes	23	(79)	21	(72)
No	5	(17)	4	(14)
No response	1	(4)	4	(14)
Total	29	(100)	29	(100)

ensure that he has time to complete his own exploitation of the data, and maintain intellectual control, but also wants the recognition and collegiality that secondary analysis implies.

Conditions listed in allowing directory access to data information:

- I don't want to fill requests myself; someone else would have to handle duplication; minimal effort on my part to disseminate; if we were not responsible for maintaining and distributing files.
- There might be a charge, e.g., Xeroxing, postage.
- Directory included adequate description, price of copies, subject to listing changes by author; short description of the data used.
- Further longitudinal studies by my co-researcher and I would be done before sending data to others; after I'm completely finished studying it and working it over.
- Certainly disclaimers about interpretations made by other people would be needed.

Conditions listed in willingness to archive:

- Don't want to fill orders myself; that we not be bothered by requests;

- Depending on the amount of local effort required to do so; if I had the money to pay for high-level clerical aid in reorganizing the data somewhat.
- No cost to us; if time, money and clerical help were available.
- Very similar to ERIC present system of handling research.
- After publication of longitudinal study; after I'm through working it over.
- Approval by me prior to release to requestor; archive handles duplication and cataloging; I am given name, address, and telephone number of persons requesting use and credited in publications resulting.

People who have shared data reported their restrictions. Four respondents had placed restrictions on the use of their data, including privacy of subjects, acknowledgment in publication, copies of reports resulting from its use, and no further distribution of the data (all from one respondent). Others required that the data be returned, that data use be for purposes similar to those of the original project, and that approval be obtained from the libraries which had participated in the study.

The results from the willingness questions do not provide a clearcut preference for archive or directory service. Control of use may be more easily maintained by a listing, although the researcher is then required to handle or delegate duplication, mailing, billing for himself.

(3) Popularity of data sets: Two characteristics of the data that researchers were willing to share are also interesting: (1) "popularity" by past request, and (2) machine-readability.

Most of the named data sets were reported as never having been requested, whereas 14 data sets have been requested from 1 to over 120

times. The identification of popular data sets poses an interesting problem in acquisitions, particularly when the data sets are as diverse as is indicated in the Information Resources scope. Only three data sets were cited as having been archived (from Association of Research Libraries and ERIC) and known requests for these were few.

It can be expected that the effort to archive, i.e., preparation and duplication of the data for archiving, would work to cull non-popular items before acquisition. The same culling might not occur for a directory, and additional sets might also be acquired for listing from those researchers concerned with personal control of data sets.

(4) Machine-readability of data sets: The data sets that are available are most often not in machine-readable form. Respondents who have generated data sets indicated their willingness to share them in machine-readable and other forms.

Table 4. Format for sharing data

Format	N=31	(%)
Printed copy	15	(48)
Magnetic tape	5	(16)
Punched cards	8	(26)
Other*	3	(10)

*(Photocopy, printout, microfilm)

Data from sharers and users of data confirm the format of data is often not machine-readable.

Data archives in the social sciences typically handle only machine-readable data. The limitation has two advantages: (1) culling of

potential sets by format where major studies are machine-readable (although some archives also code major studies that are not machine-readable); and (2) the benefits of multiple use without rekeying.

The ERIC situation differs because it would be possible to incorporate hard-copy data in the document processing flow as it exists now; new systems design to incorporate machine-readable data itself would be required. The advantage of culling by format may be specious given the data currently available and current data sharing patterns.

Estimate of data sets available to ERIC/IR

Estimating the number of data sets available to ERIC on a yearly basis is subject to a number of conditions. In order to achieve an estimate for operational projections, a set of assumptions were made. How the tradeoffs identified in this discussion will affect the availability of data sets awaits actual operation of a sharing service.

Approximately 500 documents can be identified yearly that might have data sets of use to others. From our survey results, approximately 25% of the authors did not report having generated data (partially due to an attempt to be inclusive in selecting the documents). If the same error rate can be assumed in ERIC document screening, about 375 data sets would potentially be available for the yearly ERIC/IR document output. Researchers fail to make data available because they have moved, or cannot be bothered. Our response rate of 50% can be used as a response rate for initial querying of availability. The result would be the linking of the ERIC/IR with 175-200 data set generators; refusal to release or contribute might further limit those available to approximately 150, or less for machine-readable data only. No distinction can be made in the behavior required to list as

opposed to archive, although some loss might be expected in the archive mode. Of the available data sets, approximately 40% might be in machine-readable form. Further reduction or expansion could be achieved through ERIC acquisition policies.

In summary, data sets on diverse topics and in a variety of formats are generated, available, and have been shared informally. The implementation of a data sharing service by ERIC/IR has unclear impact on the current system. There is, however, a willingness to share through a service, even of unknown quality. And there are data sets, approximately 150 a year. The challenge in providing a data service is to identify the appropriate data sets and to clear them for data sharing.

DEMAND FOR DATA SETS: FINDINGS

Data archives provide services primarily to researchers who are seeking data for secondary analysis. The audience for data for secondary analysis can be assumed to be small in contrast to the audience for ERIC bibliographic services. An intent of the survey was to query the need for data within the researcher segment of the ERIC/IR audience. Because data access also benefits instruction in research and doctoral education, the group of researchers was expanded to include doctoral advisors and instructors of research methodology.

Inadequacies of current data sharing system

Results indicate there is a demand for research data for secondary analysis within the scope of the ERIC/IR Clearinghouse. Our respondents cited data sets sought but unobtained and data sets borrowed. Data sets are used and the activity of data sharing is more prevalent than expected. The interest in data sharing described was for single data sets. Access to multiple data sets for cumulation was a benefit our respondents did not cite. Finally, the suggestion of a data sharing service was positively received, and the respondents indicated some preferences in the mode of service.

Responses to our first query indicated that for some researchers the idea of a data sharing service was not new:

- "I think you're onto a very good idea."
- "A good idea."
- "We need an archive of the treatment used in media research."
- "Strongly urge you to proceed."
- "I do believe the profession needs such a capability."

- "Bless you. And I talked about the need for such a service in 1974."
- "Good idea."
- "Sounds great to me."

On the initial postcard, 36% of the respondents indicated having wanted data that they did not obtain. The cases in phase 2 gave substance to the problems of acquiring data:

- Reasons for not receiving data when requested (all requests were to originators):
  - data lost
  - did not acknowledge letter; no response; no response to 3 letters in 6 months
  - owner had a sense of propriety
- Reasons for not requesting data:
  - did not know who to contact; insufficient citation
  - question of propriety
  - cost, couldn't budget
  - generated data from own resources.

One person who said he had been told the data had been lost noted: "In my case, a data archive would have been the answer...the degree of 'lostness' of the data may be questionable."

The most frequently mentioned source for becoming aware of the data was through publications (n = 7). Four others came to know about it via contact with the originator. People who were able to obtain data, in contrast, became aware of its existence through contact with the originator. Apparently those researchers with a personal relationship with originators have a better chance of obtaining the desired data. One respondent summed it up: "People are often wary of such requests; they sometimes suspect professional association (paranoia)."

Data sharing services formalize the data acquiring activity. Bibliographic control and access information together serve first those

researchers who felt they had no access. Second, source information or archive duplication overcomes the problems of obtaining data. Miller (1976) notes that archives in the social sciences created access for more researchers since access then does not rely on a system of personal contact. At least some of our respondents failed to achieve access through the informal network.

The personal contact system has worked for 60% of our respondents. Awareness of data came primarily through personal contact or referral (a small proportion identified data sets through publications and other formal channels). Data was obtained from the originator.

Characteristics of data sets shared

The nature of data wanted and data obtained is similar to that identified as available (previous section). The topics are diverse: research modes include surveys of individuals and institutions and experimental studies (table 5). Again, the data are only sometimes shared in machine-readable form.

Table 5. Listing of data sets shared*

Application of DeProspo's Performance Measures
CASC Library Survey Data
Circulation Count 3 years
Cooperative Education Project
Costs of Books Shipped
Drug Use Study
Helping Minority Nursing Students
Information on New England Libraries before 1855
Libgis/Hegis Data
Number of Books Shipped
Oklahoma Census
Public Library Statistics
Relative Clause Tasks
Role of Community College in Minority Education
Scientific Information Use
Southern Public Library Statistics
Statistics: MECC
Studies on Weeding
Why Minority Students Frequently Fail

* Respondents reported having obtained these data sets for secondary analysis. Researchers used unique names for data sets; those with substantive titles are listed here.

A major finding of our survey was that although the data was not machine-readable, the formats of data when shared were considered adequate by the users. The data typically arrived in a usable condition (90%) and information about the data was available through codebooks, articles, questionnaires, or conversation with the originator. Data use was perceived as not requiring major effort and well worth the effort. Our respondents reported they would do it again.

Uses and benefits of data sharing

The benefits of data sharing come in the uses of the data shared. Our respondents used data for a variety of purposes and achieved a variety of products:

Table 6. Purposes for secondary use of data sets

	N=28
Teaching	4
Secondary analysis	5
Information about another program, library, etc.	3
Resource	5
Comparison with own work	10
Other (periodical publication)	1

Table 7. Results of use of shared data sets

	N=28
Publication	7
Teaching unit	3
Background, comparison	9
Communication with originator	3
Ph.D. dissertation	1
Other (degree, testing)	2

Class-related use of research data can allow students to learn statistical techniques from a hands-on approach and may have a direct impact on the quality of research in the field. Ten of the people (N = 26) already use data sets for course activities or assignments. Use for instruction follows the pattern of data use generally.

The data sets that are currently being used are primarily those generated within the department or borrowed through personal contact with researchers outside the school. One instructor reported use of NORC data available through several data archive services in the social sciences. Most instructors (75%) responded that they would use data

sets for instruction if they were generally available. Several people commented that their use would depend on cost considerations, convenience, documentation accompanying data sets.

Since doctoral students are researchers with special requirements, constraints, and a long-term professional life, we wanted to ascertain whether advisors would allow their students to use borrowed data in their program (14 of 18 respondents would). Doctoral uses for data would include: data use for learning statistical techniques (13), for background research (11), reliability checks on their own work (10), and primary data for dissertation work (8). Again support for a data service was expressed.

Preferences and implications for ERIC/IR

Consideration of demand for a service differs from demand for data in that the implementation of a formal service must be hypothesized. As noted in the section on levels of service, an array of services might be appropriate for ERIC/IR. Our respondents to the full questionnaire indicated their preferences for both themselves and for the field. They also distinguished clearly between the current informal system and a data access service. They prefer a new service. It is interesting to note the distinction even for those whose past attempts to obtain data were successful.

Within the potential range of services, no clear preference was indicated. The lack of difference in preference for archive and directory is particularly noteworthy for system design. The following points from the survey are pertinent:

- The effort to share data is perceived to be worthwhile in the context of personal contact.
- For some sharers and users, personal contact to control usage is essential; for others, personal contact has been a side benefit of data sharing.

One conclusion then is that a system that maintains the opportunity for personal contact would be appropriate. But

- Cases of failure were attributable to no access to information about data or about its source and availability.
- A system of personal contact for information about data is perceived as less useful than a formal system.

Our conclusion then is that a service of information about data at least is needed and wanted.

Establishment of an archive for data delivery and quality control may not be necessary.

- Cases of success were perceived as worthwhile and repeatable.
- The variability in data sets used was acceptable.

Also the imposition of an archive service may reduce the collegial benefits of data sharing through informal channels or a directory and would exclude data generators concerned with data control.

To summarize, there is a system of personal contact that appears to provide adequate access and data delivery in many cases. Moreover, for successful seekers, the uneven nature of the data and the effort required to use it are acceptable. The major criticism of the current system is that it is uneven: some seekers fail to find and obtain wanted data sets.

ORGANIZATION AND DESCRIPTION OF DATA SETS IN ERIC/IR

Description of data sets in Information Resources is considered in this section. Precedents and requirements identified in social science archiving are discussed first. The amenability of the current ERIC resume format for this type of information is then considered.

Requirements and precedents

The objectives in creating a description of a data set are twofold:

1. to ensure adequate unique identification for inventory purposes and for selection among data sets when appropriate data sets are known, and
2. to ensure retrieval of appropriate data sets when these are unknown.

The literature suggests that the first objective is met by the data elements listed below, with back-up information available through references and codebooks.

Table 8. Data elements required for inventory purposes

1. Authorship, or responsibility for origination of the data.
2. Title of the data set; subtitles for distinguishing data subsets of larger sets.
3. Organizational propriety and/or corporate funding of the data collection efforts.
4. Dates: of data collection and of data set publication.
5. Availability: particularly if this differs from 1 or 3, source and price information; constraints on use.
6. Physical description: storage medium, number of cases; number of variables; description of subfiles if these exist.
7. Contents information: objectives of study, major variables or classes of variables, sample design and data types (e.g., longitudinal, from records or survey, unit of analysis, subcategories that affect case ordering, such as geographical.

8. Unique identifiers: source inventory numbers, archive accession numbers, or other control numbers.

   (Two other information fields are particularly helpful in the inventory record for selection among data sets.)

9. References: listings of publications in which data collection and analysis were described (see ICPSR sample in Appendix IV).

10. Documentation listing: availability of codebooks and questionnaires with source and price if these are variable.

The format of records varies across archives and data listings although authorship, organizational propriety, or title of the data sets are the typical main entries. (See Appendix IV)

Indexing for retrieval of data sets when a particular set is not known shows more variability in current social science practice. Indexing individual survey items has been the posited ideal (Nasatir, 1973, Phillipson, 1968). The Roper Center indexes by item both by keywords in the particular item and by controlled vocabulary. The user can expand a study by using more than one data set, even where the particular item was not central in the original data collection. With keywords, the user can seek items comparable in operationalization that may have been used for conceptually different research purposes, whereas through the controlled vocabulary, conceptual similarities are linked.

Nasatir (1973) argues that a first step to this method of indexing is the machine-readable codebook. On-line retrieval of the index file is also essential.

Other systems have used controlled vocabularies and classification schemes to index the major topics of the study, the populations, particular instruments, and major variable classes. The overall indexing method requires preparation of an abstract, classification, and indexing of the data set as a whole.

The item-by-item indexing technique has the advantages of:

1. access to all studies which include particular items or variables,
2. full exploitation of the data sets directly through the index,
3. potential increased use of each data set.

The advantages of the item-by-item indexing accrue to the service whose holdings are homogenous in methodology and topic area, whose data sets are of relatively even quality and whose budget can support the operational costs of inputting codebooks (if this is not already done), of indexing, and of maintaining a retrieval system. Item-by-item retrieval works where queries of possible data sets are made to the service; hence, it also appears to require established credibility in the field which is served.

The over-all indexing system has the following advantages:

1. simpler input requirements for preparing a new acquisition for access,
2. fewer entries per data set (important in cumulative printed indices),
3. no requirement for machine-readable codebooks and machine retrieval.

Overall indexing provides an entry vocabulary only. It requires adequate back-up information on operationalization of major variables and auxiliary variables. In this situation the user is required to anticipate conceptual topics under which items of interest might appear. There is less probability of retrieving a relevant item across data sets. It is, essentially, less effective in handling the second type of query--that is, when particular data sets are not known. As a system, it appears adequate when recurring variables are not frequent, i.e., when the data sets are heterogenous, and when primary use is not for the purpose of cross-study comparison on auxiliary variables.

Nasatir (1973) has summed up the issue on level of indexing:

> The question is, should the archive on cataloguing place great emphasis on detailed cataloguing or is the burden more appropriately placed on the user to develop more sophisticated searching strategies. This is not a question that has yet been resolved to a degree satisfactory both to archives and to users. (p. 42)

Choice of indexing level for Information Resources is affected by two distinctions from social science research archives. The research studies are a heterogenous mix. Hence, the potential for cross-study comparison initially is lower. This contrasts to economic and political surveys that vary primarily by time, geographical region, or population, but that contain similar research objectives. Also the data sets in Information Resources are less likely to be in machine-readable form (evidence in survey results). The more sophisticated storage of codebooks in machine-readable form is highly unlikely. Also, those archives offering item-by-item indexing have matured through increased holdings and experience in handling user queries; their design is based both on volume and on anticipation of the user language. For these reasons, and others indicated in the cost-benefit section, the item-by-item indexing is considered an unrealistic ideal with spurious benefits for the field of Information Resources. Adequate description and indexing of the data set as a whole is the recommended approach. This approach might be reconsidered if a volume of homogenous work or a demand for cross-study comparison arises as the service has matured.

The ERIC format

The ERIC format is amenable to the inclusion of descriptions of data sets. Its amenability is partly a function of its design to handle research materials and subsequent specification of required fields of information, and partly a function of its indexing and abstracting policies.

The requirements of description for data sets, and the appropriate ERIC data fields are listed below. (The full ERIC resume form appears in Appendix III.)

Table 9. The applicability of ERIC description to data sets

Requirement for data set description	ERIC field
1. Authorship	Author
2. Data set title	Title
3. Organizational propriety/ corporate funding	Source code and Institution Sponsoring agency code and Sponsor (funding work, other than that listed in source.) Contract numbers Grant number Bureau number (reserved for OE project control numbers)
4. Dates of collection; publication	No field; possible to enter in abstract. Publication date
5. Availability	Level: copyright release; level of availability Availability, including source address, control number, and price
6. Physical description (storage medium, etc.)	Note, including pagination, other descriptive information
7. Contents information	Abstract Descriptors and Identifiers
8. Unique identifiers	Clearinghouse control number Source control number
9. References	Note
10. Documentation listing	Abstract

There are two alternative treatments for data sets possible within the ERIC practices: (1) treatment as a supplement to the report listed in ERIC/RIE or CIJE, and (2) treatment as a separate document. Both alternatives assume that indexing at the data set level is adequate.

Alternative 1: Treatment as a supplement to the resumes in ERIC indexes (RIE and CIJE)

This treatment adds the data set information to the record generated by ERIC for the research document. The technique requires that the source document is processed by ERIC. It also implies concurrent processing of the document and the data set. Also, it requires that the indexing for the document provide adequate access to the data set. Several fields of information in the ERIC resume would need to be altered:

Availability

The ERIC Processing Manual (1974) takes into account "supporting material" in the availability field:

> Unpublished supporting documentation, not included with the report in hand, or separate audiovisual components associated with the document, should generally only be referred to if some mode of access can be provided the user, and the source cited is willing to assume the administrative burden involved, as in the following examples:

U1 AVAIL Copies of tests used in the evaluation are available from ERIC/CRESS, Box 3AP,
2 Las Cruces, New Mexico 88003
3

U1 AVAIL The Concordance of the Works of Schiller used as a basis for this study is
2 available in computer printout form from MLA/ERIC, 62 Fifth Ave., New York, New York
3 10011 (on loan)

U1 AVAIL The 351 color slides and audio tape described in the abstract are available from
2 the National Audiovisual Center, Washington, D. C. 20409 ($85.00)
3

In using the field, the following information is mandatory when applicable: full name of the source; complete address, including street number or post office box number; catalog, stock or order number; and price. The field does allow for some description of supporting documentation and could include those elements of data set physical description, e.g., tape or cards, duplication fee, etc.

Abstract

Guidelines for abstracting reports (research and technical) are:

> These usually require INFORMATIVE abstracts, which should contain: (1) objectives and hypotheses of the research, (2) methodology, (3) findings and conclusions, and (4) specific implications or recommendations for further research. Methodology should be summarized accurately and include subjects, variables, controls, conditions and testing. If limitations to the study are mentioned, they should be included. Background information should be used only if necessary to explain the objectives. (p. 205)

The data elements identified as meaningful information in the ERIC abstract are those cited in contents for data description in social science data archives. The distinction, however, is that the report as a whole (of which the data set is only one part) is the source of the abstract information. In those cases where the data set supports only part of the major document, the abstract would need to make the distinction.

Indexing

The ERIC Processing Manual (1974) recommends that the following elements be indexed for ERIC documents (the list constitutes a checklist to "guarantee reasonable exhaustivity", p. 235):

- Population concepts: educational level, other groupings
- Action concepts and material concepts, including methodology
- Curriculum concepts
- Document form/type

The document type or form is both a fixed field code and a source of descriptors (p. 179, p. 241). These are used to represent the forms or type of document "NOT THE SUBJECT" (p. 177).

The use of publication type field is precluded when the data set is treated as part of the document record, since it describes the document as a whole. The use of the descriptor field for part of the document is not explicitly excluded. The appropriate terms are "Statistical Data" and "Tables (Data)." A new descriptor, with unambiguous application to cases of data supporting the document at hand, such as "Data Set" or "Coded Data" might be instituted to identify the new type of data set access suggested here.

Descriptors indicating the occurrence of source data would necessarily be major descriptors. in order to ensure access through the printed index.

The major disadvantage of tagging the data set within the ERIC document resume is the level of specificity achieved this way. The treatment necessarily reflects the whole document (of which the data set is only a part). The problems are similar to those of series or multi-volumed sets, for which ERIC provides treatment either as separate items or as a set with analysis (p. 157). These techniques allow physical and content description of individual parts and linkage with the "parent" document. In those cases where the indexing of the source document deviates substantially from the data set indexing, or when the data sets are complex (with subsets, for instance) treatment by analytics should be considered. The indexing of the data set as an analytic is similar to treatment as a separate document (below).

A second disadvantage of treatment of the data set as a supplement to the ERIC document is the requirement that ERIC process the parent document

and the data set concurrently. Lag in receipt of either element requires either holding the first document or separate treatment.

Against these disadvantages are the advantages of limited increase of the number of ERIC entries; automatic linking to source documents; and minimal additional handling in processing, while still providing access (through descriptors) to data sets.

Alternative 2: Treatment as a separate item in ERIC

In this treatment, the data set flows through the acquisition and control process as a separate entity. The concern here is primarily with the amenability of the ERIC resume fields when the data set is the document for control.

Table 9 indicates that the essential elements are available in the ERIC resume. A few fields require special consideration, however, when no "hard copy document" is processed.

Copyright, reproduction release, level of availability

Data sets, like documents, may be copyrighted. This does not necessarily preclude entry if (1) release is obtained for ERIC reproduction or (2) an alternative source is cited. Reproduction of data, in forms other than hard copy (Level I) and microfiche (Level I or II) requires a different availability code, or a linkage that identifies the document level with added information in Notes and Availability fields.

Publication type

A new code for data sets is advantageous although N (Numerical and statistical tables; quantitative data and analysis) may be a broader designation under which data sets might be subsumed. Specific designation of data sets would allow retrieval in coordination with a topic area,

where N implies a broader range of document types. For those users specifically seeking data for secondary analysis, a new designation is preferable. Publication type in Descriptors again would best be augmented by an appropriate designation "Data Set" or "Coded Data" to allow for precise searches where the vocabulary now includes only hard copy forms as "Statistical Data" and "Tables (Data)."

Abstract and indexing terms

Again, the descriptor checklist and guidelines for abstracts of research reports serve well for data sets. Separate treatment of the data set improves the representativeness in these fields. The generation of the abstract requires some readable information; as in the social sciences, primary sources should include the codebook, questionnaire, and descriptions of research methodology. Selection of index terms, without the source report, requires classification of variables; Nasatir (1973) notes that classification requires high-level indexers, since concepts often do not appear in the operationalized items. Clearinghouse specialization with indexers trained in the subject area is a major advantage for this sensitive indexing task.

The primary advantage of treating a data set as a separate entry into the ERIC system is the level of specificity achieved throughout the record. Particularly, the record is more specific in the physical description, document type, availability and content fields, and in the entry points. Other advantages are (1) independence from the allied reports in the work flow and (2) potential to control data sets where no report exists in the ERIC system.

Controlling data sets requires a departure from ERIC guidelines on document delivery. The ideal within ERIC is the Level I document available

through ERIC in hard copy or fiche. Availability from ERIC for data requires a different delivery system.

Moreover, in those cases where the report or the codebook is not controlled, the possibility exists that a user will retrieve a citation to data without the ability to scan hard copy indicators of its contents (other than the resume). This condition can be avoided by an acquisition policy that requires either the report or the codebook and a policy of linking these in the records. In that event, the user has access to supporting information that describes the data set in depth; then acquisition of the data set can be initiated. The technique is similar to that of ICPSR where the user first accesses a summary record; then obtains the codebook and references; then, if satisfied as to the value of the data, acquires the data set.

The disadvantages of treatment as a separate record are the required modification of the document level codes, increased number of records, and a requirement to link the data set record with related reports in the ERIC data base.

Summary

The ERIC control system is amenable to inclusion of data sets if data sets are indexed as a whole (in contrast to item-by-item or variable-by-variable indexing). Two alternative treatments are possible: treatment as a supplement to the report controlled by ERIC, or treatment as a separate document. The first may compromise specificity in the physical description and in indexing when data only supports part of the report. In such cases, treatment by analytics provides a solution. The alternative,

treatment as a separate document, improves specificity but still requires links to the report when controlled by ERIC. The second technique can be enhanced by an acquisitions policy requiring a codebook in hard copy form. Either alternative would be enhanced by a new descriptor for the document type, "Data Sets."

COST EVALUATION OF ALTERNATIVE SERVICES

The following alternatives have been identified as levels of service for improved data exchange:

1. No new service instituted by ERIC: Continue the current system of informal exchange.

2. Listing of data sets in ERIC Resources in Education: This implies an on-going listing, and indexing through RIE but with responsibility for duplication with the data originator.

3. Data directory: An ERIC/IR bibliography or directory of data in the field that researchers offer for duplication, listing any constraints on use, etc.; again with duplication responsibilities for the originator.

4. Archive: An ERIC service center that gathers, preserves and disseminates data for secondary analysis.

5. Archive with added service of data manipulation in preparing subsets of data on demand.

6. Archive with added service of statistical summaries or analysis and possibly cross-study comparisons.

7. Archive with statistical consulting on the use of data sets.

8. Archive with the development of instructional data sets and instructional modules for data analysis.

To summarize the results of the cost evaluation, as one proceeds from top to bottom of the list above, the services are increasingly expensive. Costs increase dramatically between those services which provide information about data sets and those services which provide the data sets. Results of the survey indicate little difference in the perceived usefulness of services 2 through 8.

Each alternative has cost factors associated with it. These factors are of two types, incremental system costs for the added service level, and user costs. User costs will be considered separately.

Specific costs for each alternative are considered below. In each case the procedure was to determine the factors which would account for a

system's incremental cost to ERIC if the system was implemented. Then, where possible, to estimate the costs that these new or increased operations would entail. Cost information was derived from two types of sources: ERIC processing costs for report literature, and costs reported from archives in the social sciences. All costs given are direct costs.

1. No change: the current system of personal contact.

   There are no incremental costs to ERIC/IR.

2. Listing in RIE:

   The cost factors involved in RIE listing are the identification of reports from which shareable data are available, and the modification of the RIE or CIJE records to note the availability of data.

   Identification requires the modification of the ERIC input procedure. When an indexer suspects that a report may have shareable accompanying data the document would be pulled out of the normal input flow after the resume is completed.

   Special processing consists of sending the author a form letter asking if appropriate data exists, if the author would permit its availability to be noted, and at least detailed source information for the data. The reports may be delayed from input by from one to six weeks (a maximum response time should be set by the Clearinghouse which balances expected response time with document throughput time requirements). If the author denies permission (or the existence of data) or if the deadline passes without response, the report is input with no resume modification. If before the deadline the author gives permission, the resume is modified to indicate availability and source information and the resume is input to RIE or CIJE as usual.

Incremental costs are (a) an increase in indexer time to spot a potential data source and pull it out of the flow, (b) construction and duplication of the form letter, (c) postage, (d) clerical time to handle the correspondence and modify the resume; well less than one hour per specially processed report. This cost is estimated at a maximum of $5 per report. Approximately 25% of RIE and 13% of CIJE reports would receive special processing, if a criterion similar to that used in the drawing of the survey sample[1] is used, which tends to err on the side of inclusion; twenty-five percent of the 1300 RIE reports input annually plus thirteen percent of the 1350 CIJE reports[2] totals 500 documents to be specially processed annually. Of these 500, based on the responses to the survey, an estimated 150 reports would receive author confirmation and approval. The total incremental cost estimate = $5 x 500 + cost of the delayed input of about 500 reports = $2500 + input delays.

3. Data set directory:

There are two options for a directory: (a) the directory is composed of descriptions taken directly from RIE and CIJE so that the same descriptions that are used for the report are used for the data set directory; (b) the data set directory has original descriptions of the data sets, described by indexers or the author. Both include the costs of RIE listing plus additional charges.

(a) For the RIE listing an availability code was marked in a machine searchable field. Production of a directory requires a computer search on this field for reports with data. This produces an output list of about

---

[1] See survey methodology appendix.
[2] Figures from performance Category Budget of the IR Clearinghouse revised 12/1/76, Attachment A, page 1.

150 report descriptions if done annually, about 38 if done quarterly. The amount of time (clerical and professional) to organize, index, and edit the list is variable, depending on the degree of sophistication desired in the directory. For estimation purposes two man-days each of professional and clerical time will be assumed. After editing the list into a directory, it would be entered as an RIE document.

Estimated cost of this directory (created annually) = (listing cost) + $10. (computer searching and printing) + $208. (16 hrs. professional x $9. per hr.* + 16 hrs. clerical at $4. per hr.*) + $52.(ERIC's computed unit cost for RIE input) = $2500. + $10. + $208. + $52 + delay = $2770 + input delay.

(b) In this system, instead of automatically selecting the report descriptions from RIE and CIJE, each data set would be treated as an independent report and its description entered into RIE. As noted in the description section, codebooks or questionnaires might be allied documents. An appropriate estimate for the unit cost for this process would be the RIE input cost, $65. per document or data set. Data entry costs would be listing costs ($2500.) + data entry costs ($65. x 150) = $12,250. + input delay. To produce a directory would then cost an additional $10. + $208. + $2. = $270 if produced annually. Total incremental costs of a listing and separate entries then is $12,520.

4. Archive

The costs of operating an archive cause the remaining alternatives to be substantially greater than listing and directory alternatives. The cost factors involved include archive start-up costs (materials and labor) and the recurring costs of obtaining, translating (reading and

*Benefits included.

writing data sets into appropriate formats for various computers and programs), cleaning and storing data sets, computer charges, computer programmer and operator salaries, documentation, processing requests, and disseminating data.

Estimates for these components are difficult to find or develop. They vary greatly with the scope and volume of the archive. Some figures and comments received from currently operating archives are informative: a list of archives providing information is appended.

Nasatir (1974) provides sample budgets for an archive (projected) in 1964 at $28,000 for a twelve month period (p. 84) and for an archive start up in 1969 at $85,000 (p. 124). The financial summary for the Inter-University Consortium listed direct costs at $552,600 for the 1975-76 year. Similarly Data Use and Access Laboratories, stressing federal data, listed expenditures of over $700,000 for the fiscal year ending January 1977 (although data access is their major service the budget includes some other contract work).

Data archives are labor intensive. One large archive noted that salaries account for 65% of the budget, data acquisition 9% and equipment, maintenance, and materials 26%. However, they did not include space, utilities, or computer costs as part of their expenses. Another archive reported having five computer programmers working for two years creating and maintaining the system.

The cost of accessioning a single reel file was estimated by one archive at $200. This figure did not include attempts to clean the data. Other costs include tape reproduction, duplication of documentation and mailing (one archive charges $300 to recover these costs); and storage (tapes at $10 per reel, storage at $.03 per day per reel).

Actual costs for an ERIC archive might be figured on the observation that total direct costs might be approximately 5 times accessioning costs (for ICPSR, total costs are 10 times accessioning costs). Costs for 150 data sets (at $200 per data set) are $30,000; direct costs then exceed $150,000.

The notion that an archive might start small fails to reduce the costs substantially, as system set-up costs (salaries, computing facilities, storage, service procedures) are fixed.

Another possible cost factor is inputting data into machine-readable form. This operation requires coding, keypunching, and programming. It is not seriously considered that ERIC/IR would convert data to machine-readable form. Unfortunately, according to the survey, only about 20% of the IR related data shared in the past two years has been in machine-readable form. 45-50% of the generators would be willing to supply data in machine-readable form, but some only at cost. The added cost of conversion is particularly important if the archive provides services beyond data delivery.

5. Archives with added services

Costs for these services are cumulative from the baseline of the archive costs. Requirements include added programming, conversion of data sets into system requirements, amplification of documentation, and specialized personnel in data handling, statistical analysis, and finally, instruction.

User costs

Data sharers and borrowers perceive no costs in data sharing. While some data generators complain about the inconvenience of preparing data

and would like an archive to handle requests and dissemination, few if any see preparation and dissemination as a cost factor. The time and money are seen to come from nowhere. Similarly, the perceived costs of obtaining data are minimal, again being "stolen" time and money.

The impact of a data sharing service might change this perception and more realistic costs might be applied. The directory and listing services include researcher time costs for indicating willingness and describing the data sets, and variable costs for duplicating and mailing on demand. In some cases preparation for use by another researcher might be an incurred cost.

Preparation for archiving is the primary cost of the archive service level (a one-time cost). This is a perceived cost to the researcher, without the benefit of personal contact. The costs for the generator of a popular data set would be lower at this service level.

For each level, there are also opportunity costs of not having the benefits of the higher service levels.

CONCLUSIONS AND RECOMMENDATIONS

This section brings together the conclusions of the sections and reports the recommendations of the feasibility study group:

1. A formal data sharing system should be implemented.
2. This system would be organized within ERIC/IR.
3. The appropriate level of service would be inclusion in the ERIC indexes and the publication of an annual directory by the Clearinghouse.
4. An archive is not recommended at this time; however, the eventual establishment of an archive for a small set of data might be considered.

Data sharing by the Information Resources Clearinghouse audience occurs now. Secondary use has benefited the individual researcher in the following ways:

- economy by reducing data collection duplication
- access to data for comparison with own work
- support of research through background information.

Data sharing also benefits the data generator, first by opening his work to peer review and re-analysis, and also by enhancing the informal communication among scholars.

Data sharing has contributed to the growth of knowledge in other ways also, through:

- publications and dissertations
- use in teaching for new professionals
- increased communication among researchers.

These benefits echo those identified in social science archiving for both individual scholars and for the growth of knowledge in a field.

But data sharing in Information Resources is unsystematic. The respondents to our survey who have personal contacts and hence access to

data enjoy the benefits of sharing, but those without appropriate contacts for a given research problem fail to achieve data sharing. Miller notes that systematic sharing through archives broadens the resources of all scholars, in a sense equalizing the opportunities to achieve new knowledge. A data sharing service improves the productivity of a field by systematizing access.

Researchers in Information Resources would prefer a formal data sharing service. A service first would provide standard information about data sets with the potential for broader coverage to the individual researcher. Second, a service would identify and possibly disseminate available data. These alerting functions are missed in the Information Resources fields. An archive would also provide efficiencies for the researcher with a popular data set. Our respondents recognized the advantages of a service to the field in addition to personal advantages in stating their preferences. Data sharing has potential benefits so far unavailable to the fields, but desired.

The ERIC Information Resources Clearinghouse is the appropriate service center for data sharing. First, it is used by the target audience for bibliographic searching and for report dissemination. Second, generators of data have expressed a willingness to list or archive their data through ERIC. Third, the system of control for report literature is appropriate for data sets.

The basic question for system design is: at what levels of service will benefits be worth the costs? The appropriate service levels within ERIC considered here are:

1. Directory: a bibliography of listing of data sets that researchers have agreed to duplicate for secondary analysis. The listing includes source information, data set description, and constraints on use; the researcher is responsible for duplication.

2. Archive: a service center which gathers, preserves, and disseminates data for secondary analysis.

The benefits of a directory service (including RIE listing and an annual directory) are:

- systematic coverage of the research data in the Information Resources scope
- awareness service to available data sets
- source information for access
- subject indexing
- maintenance of generator-seeker communication, with consequent personal control for the generator and personal contact benefits.

Additional benefits with archive service, but losing the personal contact benefit are:

- data delivery
- continued access when the generator has moved or lost interest
- quality control
- efficiencies for the generator of the popular data set and for the seeker.

The major findings that have affected our recommendation for a directory are that:

- respondents to the survey perceived a difference between the current informal system and a formal system and prefer the new system,
- but make no great distinction between a listing and a directory service in contrast to an archive service.
- respondents generally express willingness to list or archive data.
- much of the data available and shared is not machine-readable, a criteria of archives.
- the costs of a listing and a directory are about $3,000 for a directory for the source document, or about $12,500 annually for separate listings, for the current literature covered by the Information Resources Clearinghouse.

- the costs for an archive exceed $100,000 per year (although the costs might vary as data in the IR scope differs substantially from the data in social science archives).

The implication of providing data directory service is the loss of archive benefits:

1. The research seeker is required to contact the generator or his designated source in the directory mode. This enhances the possibility of personal contact but has the disadvantage of possible failure when the generator no longer holds a copy of the data, when the generator has moved, or when he fails to respond. Also, it requires more effort with less probability of success for the requester. The Clearinghouse might alleviate the problems resulting from one-time entry by revising data set records on a periodic basis and by issuing data-set updates or cumulative revised directories.

2. Quality control is less possible when the service does not obtain the data set and inspect it before inclusion. Data formats, the existence of codebooks, and other documentation can vary greatly, for example. Consequently, effort to use data is more variable. The discussion of availability indicated some of the ways the opportunity for listing or archiving might affect availability: culling by effort to include, effort to share, ability to control use, requirements for machine-readability. It can be expected that the range of available data sets is greater for directory service; quality will also probably vary more.

The variability and lack of control is appropriate for the current state-of-the-art in Information Resources. The obvious exclusions, e.g., surveys only, machine-readable only, are not those that have affected data sharing by our respondents. First, directory to non-machine-readable data sets is appropriate since users in the past acquired

data in other forms and were satisfied. Also, the range of listings allows a laissez faire sharing to operate with the potential for monitoring, so that over time, acceptable quality and topic criteria can be derived from sharing behavior. In short, directory service can lead to a better statement of availability and demand, so that archive service for the best might eventually be possible. The heterogeneous mix of data sets now generated and shared and awareness through informal contact preclude systematic quality acquisition for archiving based on the choices within our sample.

3. Eventually, the establishment of an archive might be considered for a small set of data. Two concurrent efforts should be taken into account: the National Center for Educational Statistics holds some library and educational surveys available to the research community, and the National Technical Information Service may eventually archive some data tapes in Information Science (a study by Data Utilization and Access Laboratories is in process). An alternative mode to archiving that is gaining popularity in social sciences is to deposit data in local archives (based on university campuses) with combined archive listings by topic area. The directory concept for ERIC contributes to that movement.

Particular steps for the ERIC Information Resources Clearinghouse are:

1. Formulation of a process to identify likely sources of data for the Information Resources literature; the sample of RIE and CIJE here is one model.

2. Solicitation of data descriptions from sources so identified.

3. Publication of the resulting descriptions for available data sets, first in RIE, using the document resume with revision.

4. Publication of an annual directory to current data sets using the RIE record or an additional resume for the data set separately.

Revision of the process to include separate data set description or to archive some data (hard copy through ERIC or machine-readable through a data archive) might be considered again when the impact of systematic access has stabilized. Retrospective listing of popular data sets might be considered as an allied service.

Our recommendations are modest. The basic premise is that access to information about data provides for many of the basic archive benefits. The inclusion within ERIC avoids the major start-up costs of bibliographic control for data. The opportunity exists, through the provision of a new kind of access, to contribute to the continued use of data for secondary analysis, and the consequent growth of knowledge in the field.

APPENDIX I: DESCRIPTION OF SAMPLE, SURVEY, AND INSTRUMENTS

Selection of the survey sample

The survey sample was selected to reflect the breadth of potential contributors and users of a data sharing service. The target audience has three major subject interests: Instructional Technology, Library Science and Information Science. Researchers, instructors, and doctoral advisors were the three major functional categories for whom the service might be designed.

No single sampling frame provided the breadth of coverage sought. Indeed, even within subject interests, no one source of names could be identified. Consequently, the survey sample was selected from the following sources to reflect subject and functional categories important in the target audience.

Researchers: First authors of research reports contributed to the Clearinghouse (listed in RIE) or indexed in CIJE by the Clearinghouse were within the scope of interest. These authors constitute a segment of potential users of and contributors to a Clearinghouse data sharing service. The May 1975 through April 1977 issues of RIE and CIJE were scanned by (1) accession number indicating Information Resources Clearinghouse entry into the system and then by (2) indexing terms and abstracts to identify those reports and articles of a research nature that were based on numerical data (e.g., "Survey," "Questionnaire," "Experiment," "Data," etc.) Two hundred document numbers were then randomly selected from that set (600 from RIE, 400 from CIJE). Author addresses from RIE or CIJE were supplemented by addresses from Social

Sciences Citation Index. Authors for whom no addresses were found were dropped. The sample based on this procedure was 140.

Instructors of research and doctoral advisors:

1. Library Science.

   a) Instructors: identified from the Library Education Division (ALA) listing of instructors in Journal of Education for Librarianship (1976). Instructors coded as teaching research methodology were randomly selected from both ALA accredited and nonaccredited schools. Of 215, 30 were selected.

   b) Doctoral advisors: selected from the Michigan Doctoral Students Memo (Winter 1977), a listing of dissertations in progress, including advisor for each dissertation. The 212 dissertation proposals were screened to determine which were most likely to be utilizing numerical research data. Since many of the proposals involved historical or biographical work they were excluded from the sampling frame. Fifteen of the 86 dissertations were chosen randomly for the sample.

2. Information Science

   Instructors and advisors. Overlap with library science was tested through membership in the ASIS special interest group on research. It was decided to augment the sample from Library Science with schools from the ASIS listing of programs in Information Science (1971-72, Supplement 1972-73) that were not listed in the LED listing. From these, five schools were

selected and a name (usually the dean or chairman) was identified. (N = 5).

3. Instructional Technology:

a.1. Instructors or researchers: Randomly selected from the 1977 membership listing of the AERA special interest group on Instructional Technology, 20 selected.

2. Instructors: augmented by a sample of schools offering Ph.D.'s in Instructional Technology from Educational Media Yearbook, 1977, 20 out of 39 schools, addressed to the chairman or dean.

b. Doctoral advisors: Dissertations (10 from 400 listed) were identified through the ERIC listing "Doctoral research in educational media, 1969-1972," (1975), advisors then identified (when listed) in Dissertation Abstracts and addresses in the University Faculty Directory. (N = 5).

Comments

In selecting the sampling frames the following problems were considered.

(1) No single source of names could be identified that was isomorphic with the target audience. The allied fields covered by the Clearinghouse are an intersection of separate disciplines. Also traditional sampling frames, e.g., professional organization memberships, or authors in professional journals, in the separate disciplines vary in representativeness of the respective segments of the population. Sampling frames also overlapped in membership (some duplicate entries were excluded). Hence the following limitations:

(a) size of the population only can be estimated from the sampling frames; the lists used represent an underestimate of researchers (at least) in the field.

(b) the subject and function characteristics of the sample cannot be assumed to be proportional to the target audience. The sampling frames provided breadth so that all cases were covered. However, the same categories were rated by respondents on the first wave to (1) provide a check on the sample typologies and (2) provide subject functional categories for the data analysis.

(2) The lists for Instructional Technology lacked currency and specificity. Whereas for Library Science it was possible to identify teachers of research methods and advisors for the last academic year (1976-1977), lists for Instructional Technology were (a) older (1971-1972) and (b) less specific (schools only, not teachers of research methods per se), or interested in research and education but not specifically teachers. It is assumed some members of the target population were thus systematically excluded. Again, the result is to underestimate the population but in this case, the sample breadth has also been extended in an unidentified way. Identifying respondents from the Instructional Technology dissertation research and instructors in Instructional Technology inconvenienced us; creation of up-to-date listings might provide an important service to the Instructional Technology community.

Given the limitations of the sampling frames noted above, projection of a level of activity for a data sharing service to the target audience would be inaccurate. Even with the best of sampling frames, such projections lack confidence since information use projections for a new service are subject to numerous outside elements.

The samples allow description of data sharing successes and failures, characteristics of current use and willingness to share. The proportional frequency of responses is useful for this description, but it is not intended to represent an accurate projection of activity with intervention of a new service.

Survey instruments

The survey was fielded in two waves in an attempt to maximize the response rate. Each person in our sample (n = 226) was sent an introductory letter with a postage paid postcard to return. The postcard (see below) was designed to provide basic data on the fields of the respondents, their roles, and access to computers as well as determining the individuals' experiences with data sharing.

The second wave, a more detailed questionnaire, could then be tailored for each respondent. The postcard format was chosen to get respondents to answer quickly; provide basic information required with minimal respondent effort; and allow us to ask more detailed questions via the second wave questionnaire on material relevant to the individual's experience.

The second wave questionnaire was divided into seven sections. The responses on the postcard determined which of the sections each individual received. All respondents received a section on the perceived usefulness to themselves and to the field of a data archive service.

125

ERIC Data Archives Project

1. What is your area of interest:
   ____Instructional Technology
   ____Information Science
   ____Library Science

2. Are you (check all applicable) a(n)
   ____Instructor of Research Methodology
   ____Doctoral Student Advisor
   ____Researcher
   ____Administrator
   ____Other (please specify at left)

3. Have you ever (check all applicable)

   ____used someone else's data
   ____wanted to borrow someone else's data, but did not
   ____generated data which might be of interest to others
   ____shared your data through any means (informal borrowing, archives, etc.)

4. Do you have access to a computer ____yes ____no ____don't know

   If yes, does it have facilities for:
   magnetic tape____yes ____no ____don't know
   card input ____yes ____no ____don't know
   SPSS or other statistical package____yes ____no ____don't know

5. Is your data handling experience on computer

   ____extensive ____adequate ____minimal ____nonexistent

6. Please feel free to make any comments:

Thank you.

Time frame

The first wave of the survey was fielded the beginning of June, 1977. We sent the 226 letters with enclosed postage paid postcards bulk rate at that time. When the initial postcard response was returned to us, the individual was sent an individualized questionnaire. This usually was done within 24 hours of receipt of the postcard. Throughout the survey, our aim was to save the respondent's time. This was done through the two-stage survey, by including postage paid return for both waves, and by getting back to the respondent with wave 2 as soon as possible.

After one month, we randomly sampled the US non-returns and sent out a reminder postcard to 25 individuals in our sample. Eight of those 25 did return their postcard after this prompt.

There was also an added benefit from the prompt because of the small number of prompts sent out, they were sent first class mail. Those with bad addresses, where the person was no longer there, were returned to us. Therefore, we can estimate the number of persons in our whole sample which may have been bad addresses, at 22 of the 91 from whom we never received a response.

While the majority of respondents who returned the postcard did so after the first month, we felt it necessary to close off our distribution of questionnaires in order to allow time for them to be returned, analyzed, and this report written. Any individuals who returned their first wave postcard later than August 15 (ten weeks after our mailing) were not sent the questionnaires. Their responses in the postcard were tallied.

Response rate

The response rate on the postal card was 50.4%; for the questionnaire, 46.3%. The table identifies response rates for the different sample frames.

Table A.1. Response rate

	Total in Section	# Returned Postcard	% Returned Postcard	# Questionnaires Returned Completed (97 Sent)*	% Questionnaires Completed of those Sent	% Questionnaires Returned Completed of Total Sample (226)
Instructional Technology	43	24	55.86	4	16.6	9.3
Information Science	5	2	40.0	0	0	0
Library Science	48	32	68.7	15	45.4	31.2
CIJE	26	18	69.2	9	50.0	34.6
RIE	88	32	36.37	17	53.1	19.3
Foreign (pulled out of CIJE, RIE samples)	16	6	37.5	0	0	0
TOTALS	226	114*	50.4	45	46.3	19.9

*Questionnaires were not sent to seven respondents who did not need a questionnaire as they were neither doctoral advisors nor instructors in research methods and had not wanted, shared, used or generated research data. Eight of the returned postcards were received too late to send questionnaires. Two other questionnaires were returned as bad addresses in phase two.

Results

The results of our initial wave, the postcard, provided us with descriptions of the intended population in addition to data on the research sharing and generating experiences of our respondents.

The postcard allowed the respondents to check appropriate boxes on their fields (Instructional Technology, Information Science, and Library Science), professional roles and functions, experiences with data handling and access to computers, as well as provide data on whether they have generated, wanted or shared data, and borrowed some-else's data.

Table A.2. Postcard responses

By Sample	1	2	3	4	5	6	7	8	9	10	11	12	Returned Postcard
IT	22	5	3	9	15	16	15	6	13	12	21	14	24/43
IS	1	1	0	1	1	2	1	0	1	0	2	0	2/5
LS	6	13	26	27	16	21	6	6	22	9	20	21	32/48
CIJE	7	7	10	7	6	14	9	5	7	5	14	11	18/26
RIE	10	13	15	7	9	22	13	9	17	12	25	21	32/88
FOR	6	3	0	2	3	6	1	1	3	3	5	2	6/16
TOTAL	52	42	54	53	50	81	45	26	63	41	87	69	
Percent for each item	45.6	37	47	46	44	71	39.4	22	55	36	76	60	

1. Instructional technology
2. Information science
3. Library science
4. Instructor in research methods
5. Doctoral advisors
6. Researchers
7. Administrators
8. Other
9. Used data
10. Wanted data
11. Generated data
12. Shared data

Table A.3. Access to computer and facilities

Type of facility	Yes N (%)	No N (%)	Don't Know N (%)	Total N (%)
Computer	105 (93.7)	5 (4.5)	2 (1.8)	112 (100)
Tape facility	97 (94.2)	1 (1.0)	5 (4.8)	103 (100)
Card facility	100 (95.2)	0 (0)	5 (4.8)	105 (100)
Statistical packages available	94 (91.2)	1 (1.0)	8 (7.8)	103 (100)

Table A.4. Experience with computers

	N	%
Extensive	13	11.5
Adequate	34	30
Minimal	50	44.3
Non-existent	16	14.2
TOTAL	113	100.0

APPENDIX II: DATA ARCHIVES FROM WHICH INFORMATION WAS RECEIVED

Data Use and Access Laboratories (DUALabs)
1601 North Kent Street
Arlington, Virginia 22209

Inter-University Consortium for Political and Social Research
Institute for Social Research
Center for Political Studies
The University of Michigan
Box 1248
Ann Arbor, Michigan 48106

National Archives and Records Service
Machine Readable Archives Division (NNR)
Washington, D. C. 20408

National Center for Education Statistics
U. S. Department of Health, Education, and Welfare
400 Maryland Avenue, S. W.
Washington, D. C. 20202

National Opinion Research Center (NORC)
University of Chicago
6030 South Ellis Avenue
Chicago, Illinois 60637

Information Center on Education
The State Education Department
Albany, New York 12224

Center for Human Resource Research
College of Administrative Science
1375 Perry Street, Suite 585
Columbus, Ohio 43201

Project TALENT
American Institutes for Research
University of Pittsburgh
135 North Bellefield Avenue
Pittsburgh, Pennsylvania 15213

University Center for International Studies
University of Pittsburgh
G-7B Mervis Hall
Pittsburgh, Pennsylvania 15260

The Roper Public Opinion Research Center, Inc.
Williams College
Williamstown, Massachusetts 01267

Data and Computation Center for the Social Sciences
The University of Wisconsin
Data and Program Library Service
Room 4451, Social Science Building
Madison, Wisconsin 53706

APPENDIX III. ERIC DOCUMENT RESUME FORM

ERIC Report Resume OCR Form

Date | Prepared By | Typewriter ID No.

☐ NEW RESUME
☐ CORRECTIONS ONLY PAGE ___OF___

Keyword	ALIGN FIRST CHARACTER IN BOX
CH	Clearinghouse Accession Number *
PA	Program Area Code
PDAT	Publication Date
LEVEL	Level of Availability *
PAGE	Pagination
AUTH	Personal Author
TITLE	Title *
INST	Institution/Corporate Author and Source Code
SPON	Sponsoring Agency(ies) and Source Code
CONT	Contract Number
GR	Grant Number
BN	Project Control Number (for Office of Education and NIE projects)
REPNO	Report Number(s)
PUBTYPE	Publication Type *
NOTE	Descriptive Note
AVAIL	Availability
JNL	Journal Citation
DESC	Descriptors *
IDEN	Identifiers
ABST	Abstract *

MANDATORY DATA FIELDS ARE IN BOLD TYPE

* Mandatory data fields

Substitution Characters

ERIC	OCR
(	((
)	))
!	??
>	.gt.
<	.lt.
{	(
}	)
#	#

DO NOT TYPE BELOW THIS LINE

EPP-100 10/78

APPENDIX IV. SAMPLE DATA SET DESCRIPTIONS

DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE/NATIONAL CENTER FOR EDUCATION STATISTICS

HIGHER EDUCATION GENERAL INFORMATION SURVEY (HEGIS) – COLLEGE AND UNIVERSITY LIBRARIES

CONTACT FOR TAPE ACCESS:	Mr. Robert A. Heintze Department of Health, Education, and Welfare National Center for Education Statistics 400 Maryland Avenue, SW. Washington, D.C. 20202 Phone: 202-245-8460
CONTACT FOR GENERAL INFORMATION:	Dr. Frank Schick Department of Health, Education, and Welfare National Center for Education Statistics 400 Maryland Avenue, SW. Washington, D.C. 20202 Phone: 202-245-8638
OBJECTIVES:	To provide specific data needed by the Office of Education and its Bureau of Libraries and Learning Resources to help identify the institutions participating in title III, Library Services and Construction Act.
DATE(S) OF DATA COLLECTION:	Fall 1973 (available mid-1976)
PERIODICITY OF DATA COLLECTION:	Biennial
METHODOLOGY:	Complete enumeration of the universe of 3,015 public and nonpublic 2- and 4-year institutions of higher education. Starting in 1972, the universe consists of all these institutions *and their branches.*
SIGNIFICANT VARIABLES:	Resources, physical facilities, staffing pattern, salaries and wages, number of employees, level of educational attainment, current expenditures, and number of learning resource centers.
COST OF THE TAPE:	$ 94 per reel (1600 BPI) – $120 per reel (800 BPI) –

Feller, B.A. Directory of Federal agency education data tapes. Washington, D.C.: U.S. Government Printing Office, 1976.

HIGHER EDUCATION GENERAL INFORMATION SURVEY (HEGIS) – COLLEGE AND UNIVERSITY LIBRARIES

DATA CHARACTERISTICS:

DATA CAN BE ACCESSED AT STATE LEVEL:	Yes
DATA CAN BE ACCESSED BELOW STATE LEVEL:	Yes
DATA ARE CURRENT:	Yes (at the date of data collection)
DATA ARE RETROSPECTIVE:	No
DATA ARE PROJECTED:	No
DATA ARE LONGITUDINAL:	Yes (except for some questions)
DATA ARE DERIVED FROM RECORDKEEPING:	Yes

AUXILIARY SERVICES:

SPECIAL TABULATIONS WILL BE MADE:	Yes; at cost; based on available manpower resources; contact Mr. Richard Heilbron, DHEW/NCES, 400 Maryland Avenue, SW., Washington, D.C. 20202 Phone: 202-245-8760
REPORT–GENERATION CAPABILITY IS AVAILABLE:	No
PUBLICATIONS ARE AVAILABLE:	Yes; single copies are available from References, Estimates, and Projections Branch, DHEW/NCES, 400 Maryland Avenue, SW., Washington, D.C. 20202 Phone: 202-245-8824

Stephen B. Withey and Robert C. Davis:
NEWS MEDIA STUDY. (SSA 3511)

Information for the study was obtained from 1919 respondents, 5 cards of data per respondent. The survey, sponsored by the National Association of Science Writers, was conducted in March and April, 1957, with an area probability sample of the adult U.S. population living in private dwelling units. Closed and open-ended questions were used to explore the role of the mass media in informing the public about developments in science and to collect data on public attitudes toward science and scientists. Information was gathered about the respondents' use of newspapers, magazines, radio and television, and about stories and programs of interest. Parallel questions for each medium ascertained the content of recently read or heard science stories and an evaluation of their form of presentation. Respondents were asked which medium was their major source of entertainment, news, and science information.

More general questions probed the respondent's opinion of the effect of science on society, of the limitations of science, and of what scientific study means. Additional questions addressed the respondent's conception of causality, views on the character and specific attributes of scientists, and the freedom of scientists to engage in any kind of research. Some 25% of the survey was repeated in a national survey in Spring, 1958, six months after the launching of Sputnik. Personal data reported included age, sex, race, education, marital status, family composition, religious preference, and group memberships. The data contain ASP and dash codes. Class IV

For questions repeated from this study see Economic Behavior Program Survey, (SSA 3632).

Related publications:

Davis, Robert C.,
The Public Impact of Science in the Mass Media: a Report on a Nationwide Survey for the National Association of Science Writers. Ann Arbor, Michigan: Institute for Social Research, 1958. 254 p. (Available from University Microfilms).

Davis, Robert C.,
"The Impact of Mass Communication," in Melvin Kranzberg and Carroll W. Pursell, Jr. (eds.), Technology in Western Civilization, Volume II. New York: Oxford University Press, 1967. pp. 323-334.

Withey, Stephen B.,
"Sputnik: Some Consequences, Expectations, and Attitudes." Ann Arbor: Institute for Social Research, 1958. 17p.

Withey, Stephen B.,
"Public Opinion About Science and Scientists," The Public Opinion Quarterly, Volume 25, Fall 1959. pp. 382-388.

Raymond Wolfinger, et al.:
AMERICA'S RADICAL RIGHT. (ICPSR 7273)

These data report the attitudes, political behavior and demographic characteristics of 307 people, out of approximately 2000, who attended and supported the "San Francisco Bay Region School of Anti-Communism" held by the Christian Anti-Communism Crusade in Oakland, California, from January 29-February 2, 1962. Information was collected by personal interviews (94) and mail questionnaires (214), with most questions taken from SRC's national election studies, Samuel Stouffer's study of attitudes about Communism, and Martin Trow's study of attitudes toward Senator Joseph McCarthy in Bennington, Vermont. Information was obtained from 307 respondents in the form of 78 variables. Class I

Related publications:

Wolfinger, Raymond E., et al.,
America's Radical Right: Politics and Ideology," in David Apter (ed.), Ideology and Discontent. New York: Free Press, 1964.

Trow, Martin,
"Small Businessmen, Political Tolerance, and Support for McCarthy," American Journal of Sociology, 64 (November, 1958).

REFERENCES

Bell, C. G. The joys and sorrows of secondary data use. In R. L. Bisco (Ed.), Data bases, computers, and the social sciences. New York: Wiley, 1970.

Bisco, R. L. (Ed.). Data bases, computers, and the social sciences. New York: Wiley, 1970.

Bisco, R. L. Social science data archives: progress and prospects. Social Science Information, February 1967, 6, 39-74.

Bisco, R. L. Social science data archives--technical considerations. Social Science Information, 1965, 4, 129-150.

Carroll, J. B. Planning and utilization of a regional data bank for educational research purposes. Cambridge, Mass.: Laboratory for Research in Instruction, Harvard University, 1965.

Campbell, H. C. A Canadian social science data clearing house. UNESCO Bulletin for Libraries, July/August 1972, 26(4), 188-192.

Catalog of information elements in the N. Y. State Education Department's consolidated data base. 1975-76. Albany, N. Y.: The University of the State of New York, The State Education Department, Information Center on Education.

Cronbach, L. J. & Snow, R. E. Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington Publishers, Inc., 1977.

Dennis, J. The relation of social science data archives to libraries and wider information networks. In J. Becker (Ed.), Proceedings of the conference on interlibrary communications and information networks. Chicago: American Library Association, 1970.

ERIC processing manual. Bethesda, Md.: ERIC Processing and Reference Facility, July 1974.

Feller, B. A. Directory of Federal agency education data tapes. Washington, D. C.: U. S. Government Printing Office, 1976.

Gerham, D., & Walker, L. A subject approach to social sciences. RQ, Winter 1975, 15(2), 132-149.

Hedberg, J. G. The technologies of Exper Sim. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, Calif., April 21, 1976.

Inter-University Consortium for Political and Social Research. Annual Report 1975-76. Ann Arbor, Mich.: Institute for Social Research, University of Michigan.

Inter-University Consortium for Political and Social Research. Guide to Resources and Services. 1976-1977. Ann Arbor, Mich.: Institute for Social Research, University of Michigan.

Miller, W. E. The less obvious functions of archiving survey research data. American Behavioral Scientist, March/April 1976, 19(4).

Nasatir, D. Data archives for the social sciences: Purposes, operations and problems. Paris: UNESCO, 1973.

Nesvold, B. A. Instructional applications of data archive resources. American Behavioral Scientist, March/April 1976, 19(4).

Peters, Paul. Describing a social science information system, networks and components. ACM SIB/SOC Bulletin, Fall/Winter 1974-1975, 6(2&3), 8.

Phillipson, M. Making fuller use of survey data. Pep Report, 1968.

Ratnofsky, A. C. & Corrallo, S. B. A view on the cost of information. Paper presented at the meeting of the American Society for Information Science, Syracuse, N. Y., May 19, 1977.

Report on Education Research (bi-weekly). Washington, D. C.: June 16, 1976.

Rokkan, S. (Ed.). Data archives for the social sciences. The Hague: Mouton, 1964.

Schoenfeld, L. F. & Cooley, W. W. Data archives as resources for research instruction and policy planning: The Project TALENT data bank. Pittsburgh, Pa.: American Institutes for Research, 1968.

Vasarhelyi, P. Unesco computerized data retrieval system for documentation in the social and human sciences. Paris: UNESCO, 1972.

Wise, L. L. The Project TALENT data bank handbook. Pittsburgh, Pa.: American Institutes for Research, March 1977.

Witt, P. W. F. Doctoral programs in instructional technology. In Media yearbook. New York: Bowker, 1977, pp. 341-355.

Zweizig, D. L. & Dervin, B. Public library use, users, and uses. In M. Voight and M. Harris (Eds.), Advances in librarianship. New York Academic Press, 1977, 7, 232-253.